Angol-magyar többszavas kifejezések szótárának automatikus építése párhuzamos korpuszok segítségével

Nagy T. István[1], Vincze Veronika[2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged Árpád tér 2., e-mail: nistvan@inf.u-szeged.hu
[2]MTA-SZTE, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail: vinczev@inf.u-szeged.hu

Kivonat Jelen tanulmányunkban bemutatjuk gépi tanulási módszeren alapuló megközelítésünket, melynek segítségével félig kompozicionális szerkezetek (FX) fordításait tudjuk automatikusan megadni. A feladat nehézségét többek közt az adja, hogy a félig kompozicionális szerkezetek jelentése nem teljesen kompozicionális, vagyis azok elemeinek egyenkénti fordításával nem, vagy csak nagyon ritkán kapjuk meg az aktuális szerkezet idegen nyelvű megfelelőjét. A probléma megoldásához a SzegedParallelFX korpuszon a már korábban manuálisan annotált FX-ket kézzel megfeleltettük egymásnak. Az így létrejött korpuszon bináris osztályozó segítségével automatikusan választottuk ki a megfelelő magyar és angol nyelvű FX-párokat.
Kulcsszavak: információkinyerés, természetesnyelv-feldolgozás, szintaktikai elemzés

1. Bevezetés

A félig kompozicionális szerkezetek (FX-ek) olyan többszavas kifejezések, melyek egy főnévből és egy igéből állnak, ahol jellemzően a főnév a szemantikai fej, míg az ige csupán a szerkezet igeiségért felelős, mint például *figyelembe vesz* („take into account") vagy *támogatást nyújt* („grant support"). Korábbi munkáink során már több olyan korpuszt is bemutattunk, ahol FX-ek manuálisan vannak jelölve. Ezek közül több esetben párhuzamos korpuszokat hoztunk létre [1,2], ahol az FX-ek különböző nyelven is jelölve vannak. Ezen korpuszokon lehetőségünk nyílt olyan adatvezérelt gépi tanuló megközelítések [3,4] megvalósítására, melyek automatikusan képesek félig kompozicionális szerkezeteket azonosítani folyó szövegekben különböző nyelveken.

Mivel a félig kompozicionális szerkezetek jelentése nem teljesen kompozicionális, ezért azok elemeinek egyenkénti fordításával nem, vagy csak nagyon ritkán kapjuk meg az aktuális szerkezet idegen nyelvű megfelelőjét. Ezen szerkezetek viszonylag gyakoriak a nyelvekben, ezért megfelelő fordításuk elengedhetetlen, ám mivel szintaktikai, lexikai, szemantikai, pragmatikai vagy statisztikai szempontból idioszinkratikus tulajdonságokkal bírnak, ezért ezen szerkezetek idegen

nyelvi megfelelőinek automatikus megadása meglehetősen nehéz feladat [5]. Ezért jelen kutatásunk során kísérletet teszünk arra, hogy létrehozzunk egy olyan gépi tanuló megközelítést, mely képes a párhuzamos korpuszokon különböző nyelveken előforduló FX-ek automatikus megfeleltetésére. Ehhez a SzegedParallelFX [6] korpuszon, ahol a folyó szövegekben előforduló FX-ek már manuálisan annotálva vannak angol és magyar nyelven, manuálisan jelöltük az egy fordítási egységen belül előforduló FX-ek fordítási megfelelőit. Így például a következő fordítási egységben

> Látták, hogy bemászik az ablakon, úgyhogy nem lehetett titokban tartani.
> She was seen climbing through the window, so it couldn't be kept a secret.

a *titokban tartani* és *climbing through the window* egy negatív, míg a *titokban tartani* és *kept a secret* pedig egy pozitív példát jelöl.

Jelen munkában elsősorban az egyes FX-ek idegennyelvű FX-megfelelőit kerestük, ezért az annotálás során csupán FX-ket feleltettük meg egymásnak, nem foglalkoztunk azokkal az esetekkel, amikor egy adott szerkezet idegennyelvű fordítását egyetlen ige jelentette.

Az így létrejött korpuszon felszíni jellemzőket, valamint morfológiai, szintaktikai és lexikai információkat felhasználva tanítottuk gépi tanuló megközelítésünket, amely ezáltal képes párhuzamos korpuszokon félig kompozicionális szerkezetek idegennyelvű megfelelőinek automatikuson detektálására és így egy automatikus szótár építésére. A módszer előnyei közé tartozik továbbá, hogy amennyiben rendelkezésünkre áll egy adott nyelvű FX-azonosító rendszer, és az adott nyelvre léteznek párhuzamos korpuszok, akkor automatikusan tudunk FX-szótárakat generálni az adott nyelv és a párhuzamos korpusz többi nyelve alkotta párokra.

2. Kapcsolódó munkák

Az összetett kifejezések gépi fordító megközelítések általi automatikus fordításainak hatékonyságát vizsgáló kutatások [7] azt mutatjak, hogy számos nyelvpáron ezen szerkezetek automatikus fordítása meglehetősen nehéz feladat. Ennek megfelelően többek közt a gépi fordítórendszerek támogatása céljából jelenleg számos aktív kutatás foglalkozik az összetett kifejezések automatikus fordításával [8]. Ezen módszerek többsége [9,10] először valamilyen automatikus megközelítés segítségével azonosítja az összetett kifejezéseket eltérő nyelveken, majd a lehetséges fordítási párok kiválasztására adnak különböző megoldásokat. Jelen munkában egy hasonló elvekre épülő megközelítést mutatunk be magyar nyelvű félig kompozicionális szerkezetek angol megfelelőinek automatikus azonosítására.

3. Félig kompozicionális szerkezetek fordításainak automatikus azonosítása

Jelen munkában elődleges célunk magyar nyelvű félig kompozicionális szerkezetek angol megfelelőjének automatikus azonosítása párhuzamos korpuszokból. Vizsgálatainkat alapvetően a SzegedParalellFX [1] párhuzamos korpuszon végeztük, ahol az FX-ek magyar és angol nyelven is manuálisan jelölve vannak. Ugyanakkor méréseink elvégzéséhez még szükséges volt az egyes fordítási egységekben előforduló különböző nyelvű FX-ek manuális megfeleltetése is. Jelen munkában csak a magyar nyelvű FX-ek angol nyelvű megfelelőinek megtalálása a célunk, oly módon, hogy minden olyan magyar fordítási egységben, ahol előfordult egy manuálisan annotált FX, akkor az egység angol nyelvű megfelelőjéből a korábban már bemutatott jelöltkinyerő algoritmus [4] segítségével automatikusan kinyertük a lehetséges angol nyelvű FX-ket. Az annotátornak ezen potenciális FX-ek közül kellett kiválasztania a magyar nyelvű FX angol nyelvű megfelelőjét. A SzegedParalellFX magyar részében összesen 1377 FX van manuálisan annotálva, a hozzájuk tartozó angol nyelvű fordítási egységekben összesen 446 FX-nek találtuk meg a megfelelő fordítását, ezenkívül további 4635 egyéb lehetséges FX-et generált az automatikus jelöltkinyerő rendszer.

Mivel a megközelítésünk erősen támaszkodik az FX-ek morfoszintaktikai tulajdonságaira is, ezért szükségesnek bizonyult a párhuzamos korpusz nyelvi elemzése. Ennek során a magyar szövegek nyelvi elemzéséhez a `magyarlanc 2.0`-t [11] alkalmaztuk, míg az angol nyelvű szövegek elemzését a Stanford elemző [12] segítségével valósítottuk meg.

4. Gépi tanuló megközelítés félig kompozicionális szerkezetek fordításainak azonosítására

Az egyes FX-ek párok automatikus azonosítására egy gépi tanuló megközelítést alkalmaztunk. Ehhez alapvetően az FX-ek automatikus azonosításához korábban már ismertett [4] felszíni, morfológiai, szintaktikai és lexikai jellemzőkre támaszkodtunk, valamint új jellemzőket is definiáltunk. A korábban már ismertett, ebben a feladatban is felhasznált jellemzők a következők voltak:

- Felszíni jellemzők: a **végződés** jellemző azt vizsgálja, hogy a szerkezet főnévi tagja bizonyos bi- vagy trigramra végződik-e. Ezen jellemző alapja, hogy az FX-ek főnévi komponense igen gyakran egy igéből képzett főnév. A szerkezetet alkotó **tokenek száma** szintén jellemzőként lett felhasználva.
- Lexikai jellemzők: A **leggyakoribb ige** jellemző az FX-ek azon tulajdonságára támaszkodik, hogy a leggyakoribb igék sokszor funkcióigeként is szerepelhetnek (például *ad*, *vesz*, *hoz* stb.). Ezért az FX-jelöltek igei komponensének lemmáját vizsgáltuk, hogy az megegyezik-e az előre megadott leggyakoribb igék egyikével.
- Morfológiai jellemzők: A **szótő** jellemző alapvetően a főnévi komponens szótövét vizsgálja. Ez a jellemző az FX-ek azon már említett tulajdonságát kívánja kihasználni, hogy a félig kompozicionális szerkezetek főnévi tagja igen

gyakran egy igéből származik, ezért azt vizsgáltuk, hogy a főnév tag szótövének van-e igei elemzése. Továbbá mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért néhány magyarspecifikus morfológiaalapú jellemzőt is alkalmaztunk. Így megnéztük a magyar funkcióigék **MSD-kódját** felhasználva az ige módját (`Mood`), valamint a főnévi komponens típusát (`SubPos`), esetét (`Cas`), a birtokos számát (`NumP`), a birtokos személyét (`PerP`), valamint a birtok(olt) számát (`NumPd`).
– Szintaktikai jellemzők: korábbi kutatásaink azt mutatták [13], hogy az FX-ek igei és főnévi tagja közt csupán néhány szintaktikai osztályba tartozó él fordulhat elő, mint például alanyi vagy tárgyi. Ezen **szintaktikai osztályokat** szintén felhasználtuk jellemzőként.
– Szemantikai jellemzők: ebben az esetben is az FX azon tulajdonságát használtuk fel, hogy a főnévi tag igen gyakran egy igéből származik. Ezért a Magyar WordNetet [14] valamint a Princeton WordNet 3.1-et[1] felhasználva `tevékenység` vagy `esemény` **szemantikai jelentést** keresünk a főnévi tag felsőbb szintű hipernimái közt.

A jellemzőkészlet kialakítása során megvizsgáltuk az egyes jellemzők értékeit külön-külön a magyar FX-re és a hozzá tartozó angol FX-jelöltre, valamint megnéztük, hogy értékeik egyszerre is igazak-e az adott FX párra. Vagyis például amikor a leggyakoribb ige jellemzőt néztük az adott FX-párra, megvizsgáltuk, hogy a magyar FX igei komponense szerepel-e a magyar leggyakoribb igék közt, valamint hogy a potenciális angol szerkezet igei tagja szintén gyakori ige-e. Végül pedig megvizsgáltuk, hogy a két ige egymás fordításai-e. A fentebb ismertett, FX-ek automatikus azonosítására már korábban használt jellemzőket további attribútumokkal egeszítettük ki. Vagyis megvizsgáltuk, hogy a két szerkezet főneveinek fordításai megegyeznek-e a szótárban, illetve hogy a szerkezet főnévi tagjának van-e szintaktikai bővítménye az adott mondatban, és amennyiben igen, annak címkéjét is felvettük.

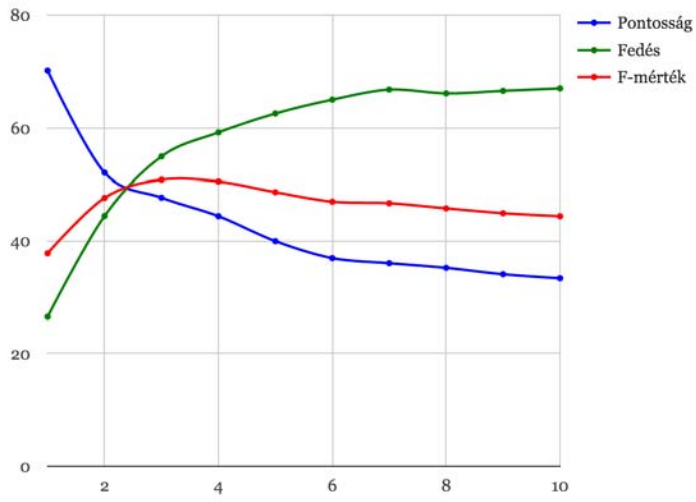
Az így létrejött tanítókorpuszon a WEKA gépi tanuló csomagban [15] található C4.5 döntési fa algoritmust implementáló J48 tanuló algoritmust alkalmaztuk. A kiértékelés során tízszeres keresztvalidációt felhasználva számítottunk pontosság, fedés és F-mérték metrikákat. Mivel a tanító korpuszon a negatív példák jelentősen felülreprezentáltak a pozitív példákhoz képest, ezért a tanítás során szükségesnek találtuk a pozitív példák felülsúlyozását. A legjobb F-mértéket eredményező súly megtalálához megvizsgáltuk módszerünk hatékonyságát különböző súlyozások mellett, melynek eredményét az 1. ábra mutatja.

Baseline megoldás szerint akkor tekintettük azonosnak egy FX-párt, amennyiben a szerkezetek főneveinek jelentése a szótár szerint megegyezik. Ezen megközelítések eredményei az 1. táblázatban láthatók.

5. Az eredmények értékelése, összegzés

Jelen munkánkban bemutattuk a gépi tanuláson alapuló rendszerünket, amely automatikusan képes magyar-angol párhuzamos korpuszokból magyar nyelvű

[1] http://wordnet.princeton.edu



1. ábra. Pozitív elemek súlyozásának hatása a gépi tanuló megközelítés hatékonyságára.

1. táblázat. Baseline, valamint a gépi tanult megközelítés eredményei

Megközelítés	**Pontosság**	**Fedés**	**F-mérték**
Baseline	73,68	15,69	25,88
Döntési fa	47,63	54,93	50,81

félig kompozicionális szerkezetek angol nyelvű megfelelőit azonosítani. Ehhez először egy manuális annotált korpuszt hoztunk létre a SzegedParallelFX korpuszon, ahol a magyar nyelvű FX-ekhez potenciális FX-eket generáltunk. Az így létrejött korpusz nem csak összetett kifejezések automatikus megfeleltetésére használható, hanem segítségével megvizsgálhatjuk, hogy mennyire hatékonyan képesek a különböző gépi fordító megközelítések folyó szövegekben az összetett kifejezéseket automatikusan fordítani.

A feladat megoldása során először a lehetséges fordítási párokat automatikusan azonosítottuk a párhuzamos szövegekben, majd gépi tanuló megközelítés segítségével választottuk ki a helyes fordításokat.

Eredményeink részletesebb vizsgálata alapján elmondhatjuk, hogy elsősorban azok az esetek jelentettek nehézséget a gépi tanulónak, amikor egy adott fordítási egységen belül az angolban és a magyarban is megtalálható volt egy FX, ezek azonban nem voltak egymás fordítási egységei, lásd pl.:

> *Háromévi várakozás után William Prichard kapitány, az Antilop gazdája, ki a déli vizekre volt indulóban, előnyös* ***ajánlatot tett*** *nekem, és én elfogadtam.*

After three years expectation that things would mend, I accepted an advantageous offer from Captain William Prichard, master of the Antelope, who was ***making a voyage*** *to the South Sea.*

Ahogy láthatjuk, a fenti esetben az *ajánlatot tett* kifejezést a rendszer megfeleltette a *making a voyage* kifejezésnek, ez azonban nem bizonyul helytállónak. A rendszernek nehézséget okozott továbbá a ritkábban előforduló igéket tartalmazó FX-ek sikeres azonosítása is (például *(nehéz) életet élnek – lead (difficult) lives*).

Ahogy azt az 1. ábra mutatja, amennyiben a tanítás során a pozitív elemek súlyát növeltük, a gépi tanuló megközelítés pontossága folyamatosan csökkent, míg a fedése növekedett. Ezen tendenciák mellett akkor kaptuk a legjobb F-mértéket, amikor a pozitív példák 3-as súlyt kaptak. Ugyanakkor a súlyozás segítségével a létrejövő automatikus szótár minőségét az alkalmazástól függően tudjuk parametrizálni. Amennyiben elsősorban pontos szótár építése a célunk, akkor a gépi tanulás során alacsonyabb súly rendelése szükséges a pozitív példákhoz, míg ha minél több lehetséges fordítási párra vagyunk kíváncsiak, akkor a pozitív példák nagyobb súlyt kívánnak a gépi tanulás során. Ugyanakkor a feladat nehézségéből fakadóan minden esetben szükséges lehet az automatikusan létrejött szótár manuális validációja.

A generált szótárakat oktatási és kutatási célra ingyenesen elérhetővé tesszük.

Hivatkozások

1. Vincze, V., Felvégi, Zs., R. Tóth, K.: Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban. In Tanács, A., Vincze, V., eds.: MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 91–101
2. Rácz, A., István Nagy, T., Vincze, V.: 4FX: Light verb constructions in a multilingual parallel corpus. Proc. of LREC (2014) 710–715
3. Nagy T., I., Vincze, V., Zsibrita, J.: Félig kompozicionális szerkezetek automatikus felismerése doménadaptációs technikák segítségével a Szeged Korpuszon. In Tanács, A., Vincze, V., eds.: MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 47–58
4. Rácz, A., Nagy T., I., Vincze, V.: 4FX: félig kompozicionális szerkezetek automatikus azonosítása többnyelvű korpuszon. In Tanács, A., Vincze, V., Varga, V., eds.: MSzNy 2014 – X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2014) 317–324
5. Sass, B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In Tanács, A., Vincze, V., eds.: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 102–110
6. Vincze, V.: Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In: Proceedings of LREC-2012, Isztambul, ELRA (2012) 2381–2388
7. Seretan, V.: Multi-word expressions in user-generated content: How many and how well translated? evidence from a post-editing experiment. In: Proceedings of the Second Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2015), Malaga, Spain (2015)

8. Monti, J., Mitkov, R., Pastor, G.C., Seretan, V.: Multi-word units in machine translation and translation technologies (2013)
9. Monti, J., Sangati, F., Arcan, M.: Multi-word expressions in a parallel bilingual spoken corpus: data annotation and initial identification results (2015) Poszter. PARSEME 5th General Meeting.
10. Wehrli, E., Villavicencio, A.: Extraction of multilingual mwes from aligned corpora (2015) Poszter. PARSEME 5th General Meeting.
11. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, INCOMA Ltd. Shoumen, BULGARIA (2013) 763–771
12. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. (2014) 55–60
13. Vincze, V., Nagy T., I., Farkas, R.: Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers. (2013) 255–261
14. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P., eds.: Proceedings of the Fourth Global WordNet Conference (GWC 2008), Szeged, Szegedi Tudományegyetem (2008) 311–320
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18